DOT/FAA/AM-99/3

Office of Aviation Medicine
Washington, D.C. 20591

Situation Awareness As a Predictor of Performance in En Route Air Traffic Controllers

Francis T. Durso
Carla A. Hackworth
Todd R. Truitt
Jerry Crutchfield
Danko Nikolic

University of Oklahoma
Norman, OK 73019

Carol A. Manning

Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, Oklahoma 73125

January 1999

Final Report

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-99/3		

4. Title and Subtitle	5. Report Date
Situation Awareness As a Predictor of Performance in En Route Air Traffic Controllers	January 1999
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Durso, F.T., Hackworth, C.A., Truitt, T.R., Crutchfield, J., Nikolic, D.[1], and Manning, C.A.[2]	

9. Performing Organization Name and Address		10. Work Unit No. (TRAIS)
[1]University of Oklahoma Department of Psychology Norman, OK 73019	[2]FAA Civil Aeromedical Institute P.O. Box 25082 Oklahoma City, OK 73125	11. Contract or Grant No.

12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aviation Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, D.C. 20591	14. Sponsoring Agency Code

15. Supplemental Notes

This research was supported by Contract #DTFA-02-93-D-93088.

16. Abstract

In this study, air traffic control instructors controlled simulated traffic while three techniques for determining situation awareness (SA) were implemented. SA was assessed using a self-report measure (SART); a query method that removed information on the plan-view display (SAGAT); a query technique that did not have a memory component (SPAM); and the detection of errors integrated into the scenarios (implicit performance). We used these measures of SA together with a measure of workload, NASA TLX, to predict two different performance measures. One performance measure was an over-the-shoulder, subjective assessment by a subject matter expert (SME). The other performance measure was a count of the number of control actions remaining to be performed at the end of the scenario. The SME evaluation was predicted by workload and the controller's appreciation of both the present and the future. The remaining-actions count (RAC) was predicted by the controller's appreciation of the future. In fact, an appreciation of the present led to poorer RAC scores: The better the participant was at answering questions about the present or the better he or she understood the present situation, the larger the number of actions remained to be performed. The results have implications for the relationships among workload, situation awareness, and performance, and suggest limitations on several of the measures currently proposed as SA techniques. The results confirm that future versus present is an important conceptual difference in air traffic control. More importantly, the results suggest that a controller who remains overly focused on the present may do so at the expense of the future.

17. Key Words	18. Distribution Statement
Air Traffic Control, Situation Awareness, Simulation, Performance Measurement	Document is available to the public through the National Technical Information Service Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	15	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Acknowledgments

This research was supported by contract DTFA-02-93-D-93088 from the FAA to Francis T. Durso.

Thanks to Scott Gronlund, Mark Rodgers, and Dave Schroeder for comments on an earlier version of this work. We would like to thank the following people for their assistance in the completion of this project. Our subject matter expert (SME), Henry Mogilka, provided invaluable support. Through his efforts, all scenarios were created and modified as needed with the help of Wayne Guthrie and Jim Ebeling. Henry scheduled instructors to participate in the study and assisted with the scheduling of the ghost pilots. Finally, Henry was very helpful with every request and evidenced consistent support. A special thanks to Rick Larson for authorizing flexibility in Henry's schedule so that he could meet our demands.

In addition, we are grateful to Dick Pollock for the use of the RTF facilities. We would like to recognize Betty (Zeke) Holmes for her efforts in arranging the necessary ghost pilots. The pilots who helped with this project were Joe Allen, Mike Evans, Russell Glazner, Scott Hughes, Bob Hutchinson, Connie Leiman, David Mitchell, and Sue Ruby.

Finally, we are very thankful to those individuals who volunteered to participate: Mark McKinney, Faith Arnell, Bob Besanceney, Skip Foster, Frank Wrisinger, A.J. Rotter, Ken Shaver, Bob Garcia, Mark Anderson, Carol Might, Mark Stemple, Paul DeBenedittis, and Peri Bennet.

SITUATION AWARENESS AS A PREDICTOR OF PERFORMANCE IN EN ROUTE AIR TRAFFIC CONTROLLERS

The issue of situation awareness (SA) has presented itself as quite a conundrum for applied investigators and basic researchers. Although SA has a number of theoretical definitions (e.g., Endsley, 1994; Fracker, 1988), most recognize SA as a cognitive construct distinct from workload (e.g., Endsley, 1993) but capable of affecting performance in a number of dynamic environments. For example, controlling air traffic is clearly a cognitive activity in a dynamic environment, and controllers recognize the value of maintaining good SA, or "the picture," as they call it.

If SA is neither performance nor workload, how can it be understood more precisely than "the picture" or more specifically than the cognitive component required to manage a changing environment? Intuitively, SA is the operator's understanding of the dynamic situation, including an understanding of the current state and likely future states of the situation. SA would include knowing the situation in which one finds himself or herself, when that situation has changed, what to do in the situation, what should follow from that situation, and how the situation relates to the operator's goals. An early, but specific, definition captures much of what is critical to SA: "the ability to envision the current and future disposition of both Red and Blue aircraft and surface threats." (Tolk & Keether, 1982 in Fracker, 1988, p. 102). Endsley's (1988a) generalization, "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (p. 97) keeps the critical aspects of Tolk and Keether's definition, while extending it beyond fighter aircraft. In both definitions, the distinction between the present and the future is highlighted.

SA is typically described as a characteristic of the operator in a particular environment. The term "environment" describes a dynamic condition in which the operator has responsibilities or goals that affect the surrounding situation. It is this goal-directed aspect of SA that highlights the importance of future events. This focus on the future helps distinguish SA from other related cognitive constructs, such as understanding or perception. Although SA includes understanding and perception, it focuses on the future more than either of the other constructs. For chess experts (Durso et al., 1995), comprehension of the current situation distinguished good players (master or intermediate) from bad players (novice) but could not differentiate between the good players. However, the ability to answer questions about the future of the game did differentiate master-level players from intermediate-level players. Presumably, good players have a better understanding of the current state than poor players, but expert players differ from intermediate players because of better representations of the future.

In this way, our understanding of SA can advance without a commitment to any particular conceptual view of SA. In the social sciences especially, operational definitions of otherwise vaguely defined constructs have often been useful starting points from which consensus conceptual definitions have emerged. In fact, for SA, several researchers have advanced our understanding by defining it operationally. Specifically, researchers have used self-report, query methods, and implicit performance measures. One straight-forward method, the Situation Awareness Rating Technique (SART; Taylor, 1990) simply asks the operator for a judgment on a number of dimensions presumably related to SA. The Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1988b) is an on-line query technique that taps an individual's recent memory of the situation. In SAGAT, information normally available to an operator is removed, and a question selected randomly from a battery of questions is presented to the operator. The more queries correctly answered, the better is the operator's SA.

In a related procedure, Durso et al. (1995), asked participants to respond to SAGAT-like queries, but all information normally available to the participant remained in view. Instead of measuring percent correct, the situation-present assessment method (SPAM) uses response latency as the primary dependent variable. Although procedurally similar to SAGAT, SPAM does differ in interesting ways from SAGAT. In addition to not requiring a memory component, SPAM acknowledges that SA may sometimes involve simply

knowing where in the environment to find a particular piece of information, rather than remembering what that piece of information is. For example, a controller need not store in memory the call sign of an aircraft, but good SA may require that he or she knows where to find the call sign, should communication with the aircraft be required. In fact, controllers are sometimes surprisingly poor at responding to SAGAT questions about information that would normally be visible to the controller (Endsley & Rodgers, 1998).

Finally, some researchers (Sarter & Woods, 1991) have argued for a procedure that assesses implicit performance. In implicit performance procedures, an error is incorporated into an otherwise typical simulation, and the operator's SA is assessed by the speed and accuracy with which the error is detected and corrected.

In the current study, we attempted to determine which of these four SA procedures, SART, SAGAT, SPAM, and implicit performance, were able to predict performance of en route air traffic controllers. For each regression model, we included a measure of workload to determine if the measures of SA supplied anything beyond this venerable construct. If SA is a viable and measurable construct, then individuals should vary in their levels of SA, and this variance should be useful in predicting performance. If it differs from workload, then SA should have predictive value above and beyond any that workload may have.

Method

Site

This study was conducted at the Radar Training Facility at the Federal Aviation Administration (FAA) Mike Monroney Aeronautical Center in Oklahoma City, Oklahoma. The facility is equipped with two radar training laboratories that allow for the simulation of en route traffic situations using the fictitious AERO Center airspace.

Participants

Twelve ATC instructors participated in the study. All participants were full performance level (FPL) controllers with an average of 18.8 years in ATC. Time as an FPL ranged from 4 to 29 years and averaged 11.6 years. The controllers had worked as instructors for an average of 7.9 years (range .17 – 38 years[1]). Participants were familiar with the airspace, but naïve to the scenarios employed.

Scenarios

All scenarios were developed in consultation with a Subject Matter Expert (SME). Five 30-minute scenarios were used. All scenarios contained a mix of general aviation, commercial, and military aircraft. Scenarios A and B were implicit performance scenarios (Sarter & Woods, 1991). Errors involving pilot readback, pilot nonconformance with ATC instructions, and data entry by the D-side were contrived by our SME and incorporated into scenarios A and B. Confederates playing the roles of the pilots and other necessary personnel were supplied cue sheets indicating what errors to perform and when to perform them. Errors were chosen to occur at varied intervals in these two scenarios. The types of errors designed to occur were those most often implicated in actual operational errors (Durso et al., 1995; Redding, 1992; Rodgers & Nye, 1993). Five errors were included in each scenario. However, one error from each scenario was not included for scoring purposes, due to difficulties in the timing of these errors and the accurate collection of data. Thus, implicit performance scores were based on four errors for each scenario. No error was scheduled to occur sooner than two minutes after the position relief briefing that began the scenario, nor with less than two minutes remaining in the scenario, nor within one minute of another error.

Scenario A was designed for an individual performing both R- and D-side functions and contained 21 aircraft: 7 arrivals, 7 departures, and 7 overflights. The four experimentally induced errors analyzed in Scenario A were 1) pilot reports discrepant altitude, 2) pilot readback error, 3) non-conforming pilot, and 4) pilot fails to acknowledge instruction.

Scenario B was designed for a R- and D- side controller team. The scenario contained a total of 29 aircraft: 7 arrivals, 10 departures, and 12 overflights. The four experimentally induced errors analyzed in Scenario B were: 1) D-side computer data entry error, 2) pilot readback error, 3) non-conforming pilot, and 4) D-side prematurely suppressed the data block. The use of a confederate, D-side controller (SME) in

[1]Participants may have worked as contract air traffic control instructors after retiring from the FAA.

scenario B allowed the introduction of data handling errors. For example, the D-side controller entered and displayed a new, but incorrect route (to Kansas City) on the radar screen.

Scenarios C, D, and E, were designed for use during testing of the three methodologies (i.e., SART, SAGAT, SPAM). These scenarios were designed to be controlled by an R-side only and were created to be approximately equal in complexity, as judged by our SME. Scenario C had 6 arrivals, 7 departures, and 7 overflights. Scenario D had 5 arrivals, 4 departures, and 11 overflights. Scenario E had 1 arrival, 10 departures, and 15 overflights. No errors were built into these scenarios.

Performance Measures

SME evaluations. The SME evaluated the controller's performance in scenarios A,C, D, and E by observing his or her behavior. The SME's participation as D-side precluded the collection of SME evaluations during scenario B. The SME used the standard on-the-job training (OJT) evaluation form (FAA Form 3120-25). The observer indicated whether a set of specific behaviors was satisfactory, unsatisfactory, or in need of improvement. Additionally, the SME wrote comments about mistakes the controllers made during the scenarios.

Remaining actions count. Following each scenario, the SME completed a remaining actions count (RAC; Vortac et al., 1993). The SME determined the control actions that remained for each flight. These actions reflect the behaviors necessary to move the flight successfully out of the controller's sector. Fewer remaining actions suggest more efficient control (e.g., Durso et al., 1998; Vortac et al., 1993). For any particular scenario, given the same starting configuration, a controller who has fewer control actions remaining at the end of a specified time is viewed as having been more efficient in moving traffic.

Workload Measure

NASA Task Load Index (TLX). Within the present experiment, we used a modified version of the NASA TLX form. The NASA TLX (Hart & Staveland, 1988) is an instrument designed to assess several dimensions of workload. These include mental demand, temporal demand, physical demand, effort, frustration, and performance. Participants were instructed to place an "x" on a line, ranging from "low" to "high" on a scale from 0 to 96 mm, reflecting their perception of their workload during each of the scenarios.

Situation Awareness Measures

Query Techniques. With the assistance of the SME, scenarios C, D, and E were examined, and six queries were designed to assess SA for each scenario. Three of the questions concerned the current situation (e.g., "Which has the lower altitude, TWA799 or AAL957?"), and three of the questions concerned a future situation (e.g., "Will DAL423 and FDX279 be traffic for each other, yes or no?"). Controllers were given a binary choice at the end of each question. With the assistance of the SME, each query, appropriate presentation times, and viable foils were selected. All questions were judged to be queries of important information by the SME.

In most respects, our implementation of SPAM and SAGAT was similar. In both, one of the six questions was presented at the appropriate time, the controller answered the question, and the response was recorded. However, the two methods differed in important respects. In SPAM, the question was presented while all information normally available remained available. The SPAM question sequence began by activating the controller's landline. Participants were informed that all phone calls would come over a single landline, and further that some of the calls would come from "CAMI center," who would query them about aspects of the situation. After the participant answered the landline, the experimenter read the question from a computer screen and initiated the timer. When the participant responded, the timer was stopped and the experimenter recorded the response.

In SAGAT, a laptop computer was placed near the participant's work area on the side of the PVD opposite the strip bay. When the time for a question occurred, the computer beeped and the scenario was frozen. Next, the participant turned immediately away from the PVD and toward the computer screen. The participant then read and answered the question by pressing the appropriate key. Once the participant responded to the question, he or she returned to the primary task of controlling traffic.

Self-report technique. The self-report method used to assess SA was a version of SART. This measurement included four scales: demand on attentional resources, supply of attentional resources, understanding, and situational awareness. During the experiment, a tone was sounded, and the scenario was frozen. The controller turned from the screen and placed an "x", for each of the four scales on a line that extended 0 to 51 mm. The time during the scenario in which each scale was presented corresponded to the time in which questions were presented in the SPAM and SAGAT conditions for that particular scenario.

Implicit Performance. In the individual version of the task (Scenario A), the participants controlled traffic with the R-side and D-side positions combined. Participants were to control traffic as they would in the field, while our SME observed to evaluate their performance. The SME measured controller performance using the OJT form. In the team version of implicit performance (Scenario B) the participants were told that they would serve as an R-side as part of an ATC team. Our SME performed in the role of the D-side operator. Trained observers recorded reaction time in seconds from the occurrence of the error to the time the participant corrected the error. The observers listened to pilot-controller communications through headphones and recorded the reaction time via a laptop computer positioned behind the participant.

Design & Procedure

Participants controlled traffic across five air traffic scenarios. Thus, a within-subjects experimental design was used. All participants first completed an informed consent form and a biographical questionnaire. Prior to each scenario, participants were given the appropriate instructions for the condition. Next, participants were directed to their control position and were provided with a position relief briefing from the SME. The briefing listed the equipment and operational conditions likely to be a factor for the air traffic positions, an overview of traffic patterns, and any problems with navigational aids. The experiment was completed in two phases with scenarios A and B in the first half, and C, D, and E in the second. Following each scenario, participants completed the TLX workload measure.

Phase one comprised the two scenarios used as tests of implicit performance. These scenarios were used to assess the participant's ability to recognize and correct errors made by pilots and other members of the controller team in a timely manner. Scenario A was always the Implicit Performance—Individual task; scenario B was always the Implicit Performance—Team task. The order of the two scenarios used for these tasks was counterbalanced. Following the completion of the first phase, participants were interviewed. They were asked about their experiences and their opinions using a post-experimental questionnaire.

The second phase of the experiment involved the participant controlling traffic while completing various SA measurement instruments. Participants controlled traffic alone. The order of the three situation awareness methodologies, SAGAT, SPAM, and SART, was counterbalanced across the remaining three scenarios: C, D, and E. Following each scenario, participants completed the modified TLX workload measure. Again, following the completion of the second phase, participants were interviewed. They were asked about their experiences and their opinions using a second post-experimental questionnaire.

Results

In all of the subsequent analyses, it is important to keep in mind that the SART, SAGAT, and SPAM tasks were counterbalanced across three scenarios. Thus, differences among these measures cannot be attributed to inherent differences in the scenarios. However, the implicit performance tasks, by their nature, demanded that specific scenarios be designed for both the individual and team versions of the implicit performance task.

All multivariate analyses used the Wilk's L test statistic. All regressions used a stepwise procedure with an a of .15; all other analyses used an a of .05. Because of the relatively small number of participants (N=12), shrinkage was addressed by reporting the adjusted R^2.

Performance Measures

Comparing scenarios. We began by comparing the five scenarios for each of the two performance measures: SME evaluations and RAC. SME evaluations using the FAA OJT form were tallied. A count of the number of less-than-satisfactory categories (i.e., "unsatisfactory" and "needs improvement") out of 27

possible was made. RAC scores were counts of the control actions remaining when the scenario was stopped.

The SME evaluations and the RAC, simply by virtue of both being performance measures, could share much in common. On the other hand, the measures certainly assess performance differently and may even focus on different aspects of the ATC task or on different components of SA. The SME evaluations are subjective, performed by an individual skilled at the task, explicitly consider a myriad of task components, and are performed throughout the task (although the final check marks may occur at the end). The RAC index is objective as argued earlier, only indirectly considers task components and in fact may focus on different task components than the SME evaluations, and is distilled to the traffic situation at the end of the scenario.

A correlation of RAC and SME evaluations across the 12 participants was conducted separately for each of the four scenarios in which both measures were taken. The two performance measures were surprisingly unrelated. The correlations were -.05 (Scenario A), -.47 (Scenario C), +.14 (Scenario D), and +.11 (Scenario E). These low, or negative correlations suggest that the information captured by the RAC differs considerably from the information reflected in the SME's evaluation. There are a number of reasons why these measures may differ, including the difference in subjectivity, the manner of data collection, and so on. However, as suggested in the later analyses, at least part of this difference is due to the fact that RAC is heavily dependent on the controller's appreciation of the future, whereas the SME evaluations depend on both present and future components.

Finally, we correlated SME evaluations from one scenario with those from another, and RACs from one scenario with those of another (see Table 1). The SME evaluation correlations tended to be quite high, with five of the six being significant. Part of the success here may lie in the fact that the SME is likely to impose additional consistency on the evaluations. The RAC intercorrelations were often more modest, with only four of ten showing any statistical significance. However, these correlations are also uniformly positive and sometimes quite substantial (e.g., r = +.87). Overall, Table 1 provides some evidence that individuals tend to maintain their relative standing in performance across the scenarios. A good controller in one scenario tended to be a good controller in the others. In general, this was true whether performance was measured by the SME or by the number of control actions remaining to be performed, although analyses suggest these two measures of performance are quite different.

Workload measures

TLX subscale scores were determined for each participant by measuring the distance from the low anchor to the participant's judgment point. With the exception of the performance subscale, the subscales of the TLX correlated highly and positively.

Table 1. Intercorrelations among the SME ratings (top) and the RAC (bottom) for the five scenarios. SME ratings could not be obtained in the Team (B) scenario.

Correlations	A	B	C	D	E
A					
SME		N/A	+.75**	+.70**	+.33
RAC		+.29	+.87**	+.19	+.53*
B					
SME			N/A	N/A	N/A
RAC			+.31	+.52*	+.62**
C					
SME				+.81**	+.63**
RAC				+.09	+.49
D					
SME					+.59**
RAC					+.15

**p < .05; * p < .10.

Intercorrelations among mental demand, physical demand, temporal demand, and effort ranged from a low of +.87 to a high of +.95. Frustration correlated less well with these factors, but the correlations were still substantial, ranging from +.42 to +.68. Thus, a controller who viewed the task as mentally demanding also viewed it as physically demanding, temporally demanding, effortful, and relatively frustrating. Performance tended to correlate negatively with the other subscales, as would be expected. The high intercorrelations among the scales suggest that, in subsequent analyses, such as the multiple regression analyses reported later, one subscale may enter the equation to the exclusion of its correlated neighbors, and it may not matter which particular subscale it is. Overall, it appears that the TLX, at least as used here as a one-time, end-of-the-scenario measure, produces two important components—workload and subjective performance.

SA Measures

SART. SART scores were determined by measuring the distance (mm) from the low anchor to the participant's judgment mark. The midpoint of each scale was 25.5, with a minimum of 0 and maximum of 51. The controllers indicated that they had an adequate supply of resources (M = 34) leading to good understanding (M = 44) and good SA (M = 45) for scenarios that they considered to be not very demanding (M = 20). The intercorrelations among the SART subscales were nonsignificant, with the exception that the SA subscale was positively and reliably correlated with understanding (r = +.88), suggesting that the controllers made little distinction between understanding and SA.

SPAM. Frequency and mean response latencies to future and present queries were computed along with the mean time to answer the landline. Participants took almost 10 seconds to answer the landline and then took another 4 seconds to answer the query. As expected, subjects were quite accurate, especially if queried about the present situation. Response latencies were comparable for present and future queries. None of the SPAM intercorrelations reached conventional levels of significance.

SAGAT. Not surprisingly, compared with SPAM, percent correct scores for SAGAT were low. There was a moderate but nonsignificant correlation between percent correct for present and future queries. If one takes the perspective that future and present queries are merely two parts of an overall SAGAT score, then the +.35 correlation represents a rather poor split-half reliability. If, instead, one takes the perspective that future and present queries capture two important, but orthogonal, dimensions to SA, the correlation then provides mild support for this thesis.

Implicit Performance. Number of errors detected and the latency to make a detection were recorded. If an error was never detected, it contributed no datum to the latency analyses. Subjects noticed as many errors when assisted by a D-side (M = 54%, 25%—100%) as when controlling traffic alone (M = 50%, 25%—75%). In addition, controllers who did relatively well in the single-staffing condition did not necessarily do well in the team-staffing condition, as indicated by the small, nonsignificant correlation (r = -.18) between the number of errors identified in the two conditions.

Predicting Remaining Actions and Expert Evaluations

These sets of analyses explored the ability of each of the SA procedures to predict the performance measures corresponding to that scenario. For example, we attempted to use SART and TLX taken during the SART scenario to predict the RAC and the SME evaluations that occurred during the SART scenario.

Given that SME and RAC were surprisingly unrelated, it is not at all obvious how models developed for predicting SME evaluations should compare to models developed for predicting RACs.

The following analyses reveal which aspects of the SA measures contribute to predicting performance above any contribution by workload. The SA contributions reflect possible differences in both subjects and scenarios. Interpretations of the regressions should not assume that the predictive value of an SA measure is due solely to, for example, differences in the controller's SA abilities. Significant SA predictors are able to detect differences in individuals, scenarios, or both.

Predicting SME Evaluations. The regression analyses for the SME evaluations appear in Table 2. SART had success predicting SME evaluations. The SART Supply subscale combined with the TLX Mental demand subscale ($p < .06$) to account for 35% of the variance in SME evaluations. Low perceived-supply and high perceived-mental-demand led to a poorer evaluation (cf., Selcon, Taylor, & Karitsas, 1991).

SAGAT also had limited success at predicting SME evaluations. The more queries about the future that a controller answered correctly, the better was his or her SME evaluation ($p < .13$), accounting for 14% of the variance.

SPAM had success in predicting SME evaluations as well. A model including the number of present questions answered correctly and the TLX Mental Demand subscale ($p < .02$) predicted 53% of the variance in the SME evaluations. As with SAGAT, the more questions answered the better evaluated was overall performance. However, in the SPAM analysis, the critical questions were present-oriented. Finally, unlike other appearances of mental demand (e.g., SART analysis), here low mental demand implied more negative comments by the SME. Because low mental demand sometimes suggests good performance and sometimes poor performance, this subjective workload component appears to be an unreliable predictor of SME evaluations.

Finally, implicit performance was also able to predict SME evaluations. In this case, temporal demand from the TLX and the number of errors detected predicted 69% of the variance ($p < .003$). Greater perceived temporal demand led to poorer performance evaluations, and the fewer errors detected, the poorer the performance evaluations.

Overall, SME evaluations were predictable by a combination of workload and SA measures. Having a high supply of resources, answering both future and present questions correctly, and detecting errors incorporated into the scenarios led to better SME evaluations.

Predicting Remaining Action Counts. Regression analyses for the RAC are summarized in Table 3. Of the SART subscales, Demand and Understanding combined to predict RAC ($p < .11$) and accounted for 27% of the variance in the remaining actions. The Demand factor is easily interpreted: The greater the overall demand perceived by the participant, the more control actions remained to be performed. However, the Understanding factor is not easily interpreted because the model indicates that Understanding and RAC are positively correlated. In other words, the more understanding professed by the controller, the more actions remained to be performed at the end of the scenario. One obvious explanation is that these controllers were not very good at reflecting on their understanding, and thus subjective measures of SA may be inappropriate in the ATC environment.

Table 2. Regression summaries predicting SME evaluations.

	Workload		SA		Adjusted R^2
	β weight	Variable	β weight	Variable	
Implicit Performance (Individual)	.0396**	Temporal demand	-1.1530**	Errors detected	.69
SART	.0099*	Mental demand	-.0270**	Supply	.35
SAGAT			-.0102*	Future queries	.14
SPAM	-.0137*	Mental demand	-.0358**	Present queries	.53

*p < .15; **p < .05

On the other hand, some of the other measures also presented similar concerns, and so we will return to an alternative interpretation of the Understanding effect after considering the other analyses.

SAGAT Future and Present queries combined with the TLX effort-subscale ($p < .004$) to account for an impressive 74% of the variance in the remaining actions. Again, part of the model is easily interpreted: The fewer future questions answered correctly, the more remaining actions were left to be performed. Also, the less perceived effort required, the better the subject performed. However, the better the participant was at answering questions about the present situation, the *greater* the number of actions remained to be performed at the end of the scenario. Because the raw correlations between the SAGAT factors and RAC were of opposite signs (ruling out a suppresser effect), we explored this further by classifying participants as poor (0 or 1 correct) or good (2 or 3 correct) on the two types of questions, present or future. This classification yielded participants who did well on both (good SA), poorly on both (poor SA), well on future but not present (future-focused style), and well on present but not future (present-focused style). The present-focused (N=4) controllers had the poorest performance with an average of 24 remaining actions; the future-focused (N=2) controllers had the best RAC performance, with an average of only 8 remaining actions.

This aspect of the SAGAT results is reminiscent of the SART understanding results and may suggest that the more one focuses on the present situation, or the more one understands (the present), then the poorer the person will be on a measure of efficiency like the RAC. Assuming "understanding" in SART is interpreted to mean understanding the present, then a similar explanation can be applied to that analysis.

SAGAT's success at predicting RAC is substantial, but it warns that some queries may be positively related to variables of interest and others may be negatively correlated. For example, imagine a battery of SAGAT queries that focused on the present; we might find that individuals who did poorly on SAGAT actually performed better on the task or actually had better SA of impending events. Thus, the current data suggest that query techniques can be improved if greater control is taken over the types of questions asked. The current study and previous work (Durso, et al., 1995) suggests that future versus present is an important difference.

Table 3. Regression summaries predicting RAC evaluations.

	Workload		SA		Adjusted
	β weight	Variable	β weight	Variable	R^2
Implicit Performance (Individual)	-.1418**	Performance			.29
Implicit Performance (Team)	No	variable	entered	the model	N/A
SART			.4585** .7026*	Demand Understanding	.27
SAGAT	-.1733**	Effort	-.1513** .2365**	Future queries Present queries	.74
SPAM			.2917*	Reaction time (future)	.13

*p < .15; **p < .05

SPAM generated a one-factor model predicting RAC. The time required to answer a query about the future ($p < .14$) accounted for 12% of the variance in RACs. Consistent with SAGAT, SPAM indicates that the slower the participants were to respond to questions about the future, the more actions remained to be performed at the end of the scenario.

Finally, for implicit performance, TLX-performance ($p < .05$) entered the model, accounting for a respectable 30% of the variance, but no implicit performance measure contributed above and beyond the controller's perceived level of performance.

Overall, the analyses of RAC scores suggest that the control actions remaining at the end of the scenario are strongly dependent on the controller's ability or tendency to consider the immediate future during that scenario. This is indicated by the results from SAGAT and SPAM, both of which suggest that controllers who answer more future queries (SAGAT) or answer them more quickly (SPAM) will have fewer remaining actions at the end of that scenario. These analyses also suggest that controllers who, instead, focus on the present situation will perform poorly on the RAC measure. Controllers who could answer more present-queries (SAGAT) or who understood the situation (SART), actually had more remaining actions at the end of the scenario. Presumably, controllers interpret the SART understanding of the situation to mean understanding of the present situation.

Predicting Implicit Performance. The design of the current study allowed us to conduct an additional analysis, namely, predicting implicit performance from the other SA measures. If SA is a unitary construct, then a good measure of SA should capture the ability of participants to detect errors. We chose to predict implicit performance from the other SA measures for a number of reasons. Most views of SA would acknowledge that the ability to detect errors is a characteristic of good situation awareness. However, the pragmatics of using implicit performance requires painstaking design of simulations, usually in consultation with a subject-matter expert, and the amount of data collected is often small, making it difficult to reach conclusions backed by any statistical power. If a simpler method of assessing SA could be developed (e.g., SART, SAGAT, SPAM), it would have a great deal of practical value. Thus, a secondary purpose of this analysis was to determine if a simple procedure could be developed within the ATC environment that could substitute for implicit performance measures.

Separate regressions were attempted for the individual and team implicit performance tasks. We expected predictability across scenarios, as is the case here, to be lower than predictability within scenarios, as was the case in the performance analyses. Nevertheless, the results were disappointing. None of the SA measures was able to predict the number of errors correctly detected in the individual case. For the team case, SART failed to produce a model capable of predicting error detection. One encouraging finding came from SAGAT which was able to predict 20% of the variance in error detection for the team situation. The only factor in the regression was the number of present-queries answered correctly. The controllers who were especially adroit at answering present questions in one scenario tended to be those who were best at detecting the errors incorporated into a different scenario ($p < .08$). SPAM produced a model with the time to answer the landline as a factor accounting for 33% of the variance. The longer the controller took to answer the landline in the SPAM condition the more errors they had detected in the earlier scenario ($p < .03$). If being present-oriented is predictive of errors, as SAGAT suggests, the longer landline times could be taken as an indication that present-oriented controllers are more reluctant to divert attention in order to answer the landline.

Discussion

The results indicate that SA measures are able to predict performance above the predictability provided by workload. Both SME and RAC measures of performance were predictable from SA measures. All SA measures were of some value in predicting SME evaluations. Both an appreciation of the present and an appreciation of the future were useful predictors of SME evaluations. Only SAGAT and SPAM, the two query methods, had any predictive value for RACs. Implicit performance supplied nothing beyond perceived workload, and SART predictions were the opposite of what one would expect.

Why did SART and implicit performance measures have difficulty predicting RAC? One possibility is that both of these SA measures focus primarily on the

current situation, ignoring the future component of SA, a component which is apparently critical to the RAC measures. The SART SA question was virtually indistinguishable to our controllers from the understanding question. In turn, the understanding question seems to have been interpreted as having understanding of the current situation. Controllers who professed a greater understanding of a particular situation did poorly on the future-oriented RAC measure. Similarly, implicit performance may lack a future component. Several facts point in this direction. First, implicit performance was unable to predict the RAC, a performance measure that was predictable by future-oriented SA measures, but not present-oriented ones. Second, predicting implicit performance depended on present-oriented factors, such as the present-queries from SAGAT. Thus, error detection may depend primarily on the present component of SA. Although an error may have consequences for the future, in some sense it is available for detection in the present. It is an interesting methodological question whether errors can be constructed that emphasize the future component of SA, or whether all errors, regardless of their future impact, are detected with equal ease "in the present." According to the current study, however, implicit performance seems present-oriented, RAC future-oriented, and SME evaluations a little of both.

Perhaps the most interesting finding was that an appreciation of the present had effects opposite of an appreciation of the future, suggesting that controllers may attend to the present at the cost of the future. Performance, as assessed by RAC measures, was not merely unaffected by the present¾ it was actually poorer when an appreciation of the present was higher. Greater understanding (SART) and correct responses about the present (SAGAT) both appeared to hurt RAC performance. Recognizing that an appreciation of the present and future can have opposite effects on performance complicates all of the measures of SA. For example, the typical procedure of randomly sampling from a pool of questions must take into consideration that a sample of questions dealing solely with the present situation can lead to a different evaluation of a system or an individual operator than would a sample of questions dealing solely with the future situation. It is not merely that future-queries and present-queries capture different components of SA, but that they may be, at least for ATC, antagonistic activities. A controller who focuses attention on the present during a particular scenario, and thus answers many such queries correctly, may well prove to be less efficient than a controller who answers fewer such queries correctly but attends more to the future.

The current study was successful in pointing to the value of an appreciation of the future. It also supplied evidence that comprehension of the current situation and projection into the future are distinguishable and important components in the SA of air traffic controllers. The present and future sometimes, however, lead to opposite effects on performance.

References

Durso, F.T., Truitt, T.R., Hackworth, C.A., Crutchfield, J.M., Nikolic, D., Moertl, P.M., Ohrt, D., &, Manning, C.A. (1995). Expertise and chess: Comparing Situation Awareness methodologies. In D. Garland & M. Endsley (Eds.), *Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness*, (pp. 295-303).

Durso, F.T., Truitt, T.R., Hackworth, C.A., Crutchfield, J.M., Ohrt, D.D., Hamic, J.M., & Manning, C.A. (1995). Factors characterizing en route operational errors: Do they tell us anything about situation awareness? In D. Garland & M. Endsley (Eds.), *Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness*, (pp. 189-195).

Durso, F.T., Truitt, T.R., Hackworth, C.A., Albright, C.A., Bleckley, M.K., & Manning, C.A. (1995). *Reduced flight progress strips in en route ATC mixed environments.* (DOT/FAA/AM-98/26). Washington, DC: Office of Aviation Medicine.

Endsley, M. (1988a). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting, 1*, 97-101. Santa Monica, CA. Human Factors Society.

Endsley, M. (1988b). Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference - NAECON* 1988, 3, 789-795. New York: Institute of Electrical and Electronics Engineers.

Endsley, M.R. (1993). Situation awareness and workload: Flip sides of the same coin. In R.S. Jensen & D. Neumeister (Eds.) *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 906-911). Columbus, OH: Department of Aviation, The Ohio State University.

Endsley, M. (1994). Situation awareness in dynamic human decision making: Theory. In R.D. Gilson, D.J. Garland, & J.M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp. 27-58). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Endsley, M.R., & Rodgers, M.D. (1998). Attention distribution and situation awareness in air traffic control. *Air Traffic Control Quarterly, 6,* 21-44.

Fracker, M.L. (1988). A theory of situation assessment: Implications for measuring situation awareness. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 102-106). Santa Monica, CA: Human Factors Society.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139-183). Amsterdam: North-Holland.

Redding, R.E. (1992). Analysis of operational errors and workload in air traffic control. In *Proceedings of the Human Factors Society 36th Annual Meeting,* (pp. 1321-1325). Santa Monica, CA: Human Factors Society.

Rodgers, M.D., & Nye, L.G. (1993). Factors associated with the severity of operational errors at air route traffic control centers. In M.D. Rodgers (Ed.), *An examination of the operational error database for air route traffic control centers* (DOT/FAA/AM-93/22, pp. 11-25). Washington, DC: Office of Aviation Medicine. Available from: National Technical Information Service, Springfield, VA 22161. Order #ADA275986.

Sarter, N.B., & Woods, D.D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology,* 1, 45-57.

Selcon, S.J., Taylor, R.M., & Koritas, E. (1991). Workload or situational awareness? TLX vs SART for aerospace systems design evaluation. *Proceedings of the Human Factors Society,* 35, 62-66.

Taylor, R.M. (1990). Situation awareness rating technique (SART): The development of a tool for aircrew systems design. In *AGARD-CP-478, Situation Awareness in Aerospace Operations* (pp. 3-1 to 3-17). Neuilly Sur Seine, France: Advisory Group for Aerospace Research & Development.

Vortac, O.U., Edwards, M.B., Fuller, D.K., & Manning, C.A. (1993). Automation and cognition in air traffic control: An empirical investigation. *Applied Cognitive Psychology,* 7, 631-651.